

~~TOPIC 5~~

Data and Data Files

Concepts of data.

Data refers to raw, unprocessed facts and figures that on their own have no meaning. Examples: numbers, names, measurements, images, sounds, survey responses, temperatures etc.

Characteristics of Data.

1. Raw ; Collected in its original form before processing.
2. Unorganized.
3. Can be qualitative or quantitative.
4. Can come from many sources.
5. Needs processing to become information.

Bit.

- A bit is the smallest unit of data a computer can store or process.
- The word bit comes from Binary Digit.
- A bit can only have two values:
 - 0 (Off / False).
 - 1 (On / True).

Representation of Data using 1 Bits.

- All types of data in IT and Statistics are represented using bits:
- Numbers.
- Letters.
- Images.
- video.
- Sound.

Example.

SHOT ON AWESOME A70 is stored as 0100001 in binary.

Byte

- A byte is a unit of digital data used in computers.
- It is made up of 8 bits.
- A byte is the standard unit for representing a single character such as:
 - a letter (A, B, C)
 - a digit (1, 2, 3)
 - a symbol @; % #
- Uses of a byte
- Measure data size (Files, Memory).
- Store characters in text.
- Represent small integers.
- Build larger units of data.

Data Types

Data types refers to the classification of data based on the kind of values it holds and how it can be processed by a computer or statistical program. They help in:

- ✓ Data storage
- ✓ Data analysis
- ✓ Choosing the right statistical methods
- ✓ Error checking in software

(A) Numerical (Quantitative) Data Types

- These represent numbers that can be measured or counted.

1. Integer

- Whole numbers (no decimal point).
Example: 5, 10, 250.
- Used to count things (e.g. Number of students).
- A** Float / Real / Double.
- Numbers with decimals.
Examples: 3.14, 2.5, -0.01.
- Used for measurements (height, weight, temperature).

(B) Character and Text (String)

Data Types

1. Character.

- A single symbol or letter.
- Example: 'A', '7', '@'

2. String / Text

Sequence of characters (words, sentences).

Example: "Kenya", "female", "BSC statistics".

Used for names, categories, labels, addresses etc.

(C) Logical / Boolean Data Type.

- Holds only two values: TRUE or FALSE.
- Represented as 1 or 0 internally (bit-level).
- Used in decision-making and conditions (e.g. Pass/Fail, yes/no).

(D) Date and Time Data Types.

Used to represent:

Dates (e.g. 2025-12-10).

Times (e.g. 10:30:00).

Datetime (Combination).



Shot on AWESOME A70

- Important in time series analysis in Statistics

Constructing Random and Sequential Data files

- A data file is a collection of related records stored on a storage device (hard disk, USB, etc).

- In IT for statistics, data files are used to store data sets for processing by software like SPSS, R, Excel, etc.

There are two main ways of organizing data in files:

i) sequential files.

ii) Random (Direct Access) files.

i) Sequential Data files

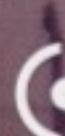
A Sequential file is a file where records are stored and accessed in a specific order, usually according to the time they were entered or a key field.

Characteristics

- Records stored in sequence (e.g. alphabetical order, date order).
- Access is linear (one by one).
- Efficient for processing large files.
- Slower when searching for a specific order record.

Constructing Sequential files.

Steps



Shot on AWESOME A70

2. Collect and enter data.
3. Store the record in the file in the same order.
4. When adding new records:
 - Often requires rewriting the entire file to keep the sequence

Advantages

- Simple to Create and Manage
- Good for batch processing (e.g., payroll, exam results).
- Uses less storage and faster to read large files sequentially.

Disadvantages

- Slow when searching for specific records
- Hard to insert or delete records
- Must start reading from the beginning.

(ii) Random (Direct Access) Data files

- A random (direct-access) file is a file where records are stored in such a way that they can be accessed directly, without reading the previous records.
- A record's location is determined by a key field using a hashing algorithm or fixed-length slots.

Characteristics

- Records can be accessed instantly.
- uses a key (e.g. ID number) to locate the record.
- More flexible and faster for updates and searches.

Requires More Storage Space than

sequential files

Constructing Random files

steps

1. Define the key field.
2. Use a hashing function.
3. Store each record.
4. If two records land in the same slot (Collision):
 - Use collision handling Methods:
 - ✓ Linear probing (next available slot).
 - ✓ Chaining (Link Records together)

Advantages

- very fast access and retrieval.
- Easy to update, insert or delete records.
- Ideal for applications requiring frequent searches.

Disadvantages

- ✓ More complex to design.
- ✓ Uses more storage space.
- ✓ Collisions may occur and must be handled.